ESTIMATION OF SYSTEM ASSEMBLY AND TEST MANUFACTURING YIELDS THROUGH PRODUCT COMPLEXITY NORMALIZATION

By

Andres Olivella Sierra
B.S. 2001, Electronic Engineering
Pontificia Universidad Javeriana

Submitted to the MIT Sloan School of Management and the Engineering Systems Division in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
AND
Master of Science in Engineering Systems

In conjunction with the Leaders for Manufacturing Program at the
Massachusetts Institute of Technology
June 2009

Signature of Author ______________
May 8, 2009
Engineering Systems Division and MIT Sloan School of Management

Certified by ______________
David E. Hardt, Ph.D. Thesis Supervisor
Ralph E. and Eloise F. Cross Professor of Mechanical Engineering
Professor of Engineering Systems

Certified by ______________
Roy E. Welsch, Ph.D. Thesis Supervisor
Professor of Statistics and Management Science and Engineering Systems,
MIT Sloan School of Management

Accepted by ______________
Nancy G. Leveson Ph.D. Professor of Engineering Systems
Chair, Engineering Systems Division Education Committee

Accepted by ______________
Debbie Berechman
Executive Director of MBA Program, MIT Sloan School of Management

This page has been intentionally left blank.

ESTIMATION OF SYSTEM ASSEMBLY AND TEST MANUFACTURING YIELDS THROUGH PRODUCT COMPLEXITY NORMALIZATION

By

Andres Olivella Sierra

Submitted to the MIT Sloan School of Management and the Engineering Systems Division on May 8, 2009 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Engineering Systems

ABSTRACT

Cisco Systems, Inc. (Cisco) has recently adopted Six Sigma as the main platform to drive quality improvements in its manufacturing operations. A key component of the improvement strategy is the ability to define appropriate manufacturing yield goals. Cisco's manufacturing operations can be divided, at a very high level, in two major steps: Printed Circuit Board Assembly (PCBA) and System Assembly and Test. The company has already deployed a global yield goal definition methodology for the PCBA operation, but the creation of a similar methodology for the System Assembly and Test operation proved difficult: Cisco lacked a universal methodology to determine the expected variation on manufacturing performance resulting from differences on product design and manufacturing processes attributes.

This thesis addresses this gap by demonstrating a methodology to relate relevant design and process attributes to the System Assembly and Test manufacturing yield performance of all products. The methodology uses statistical analysis, in particular Artificial Neural Networks, to generate a yield prediction model that achieves excellent prediction accuracy (4.8% RMS error).

Although this study was performed using Cisco Systems' product and manufacturing data, the general process outlined in this exercise should be applicable to solve similar problems in other companies and industries. The core components of the methodology outlined can be easily reproduced: 1) identify the key complexity attributes, 2) design and execute a data collection plan and 3) generate statistical models to test the validity and impact of the selected factors.

Thesis Supervisor: David E. Hardt
Title: Ralph E. and Eloise F. Cross Professor of Mechanical Engineering Professor of Engineering Systems

Thesis Supervisor: Roy E. Welsch
Title: Professor of Statistics and Management Science and Engineering Systems, MIT Sloan School of Management

This page has been intentionally left blank.

ACKNOWLEDGMENTS

I would like to thank Cisco Systems and the Leaders for Manufacturing program for providing me with the opportunity and resources to conduct this study. In particular, I would like to thank Sachin Kothawade from Cisco Systems for entrusting me with this very important project and for his constant guidance and support.

During my internship at Cisco, I had the opportunity of working with a very talented team of people. In particular I would like to thank Leslie Averbeck, CW Foong, Rosalio Florez, Thomas Dirrig, Hitesh Merchant, Joe Dickson and Susan Houlihan. Their collaboration was essential to move this project forward and working with them was a terrific learning experience.

I had the fortune of having two excellent academic advisors, Professors Roy Welsch and David Hardt. Professors Welsch and Hardt provided keen insights and advice throughout the execution of the project and their site visits and project reviews were always insightful. They are the source of many of the key ideas presented in this work.

I would also like to thank my parents, sister, and brother for encouraging me to dream and work hard. Without their kind support this experience would not have been at all possible.

Most of all, I would like to thank Lina, my wife, and Nicolás, my son: they are my inspiration and the sparks that lighten my days. Thank you for your unconditional love and the sacrifices you have made to join me in this adventure to fulfill my dream. I love you both more than anything else in this world.

Finally, I would like to thank God for continuing to bless me much more than I could have ever imagined.

This page has been intentionally left blank.

TABLE OF CONTENTS

This page has been intentionally left blank.

LIST OF FIGURES

This page has been intentionally left blank.

LIST OF EQUATIONS

This page has been intentionally left blank.

LIST OF TABLES

This page has been intentionally left blank.

GLOSSARY

ANN: Artificial Neural Network

DF: Direct Fulfillment

DPM: Defects per Million

DPMO: Defects per Million Opportunities

FPY: First Pass Yield

IP: Internet Protocol

PCBA: Printed Circuit Board Assembly

RMSE: Root Mean Square Error

RTY: Rolled Throughput Yield

VIF: Variability Inflation Factor

This page has been intentionally left blank.

1 Introduction

This thesis explores the possibility of generating a first pass yield (FPY) prediction methodology for electronic systems assembly and test operations. Having a methodology to predict FPY performance provides companies with the foundation to define manufacturing goals that reflect differences in performance due to product complexity, enabling the creation of a platform for cross-organizational learning. Learning is facilitated by the removal of the barriers that prevent comparisons between simple products and more complicated systems: since complexity is taken into account when defining each product's goals, a simple device that is not meeting its goals may be able to learn something from the very complex system that is exceeding its goals, even if the actual yield numbers are higher for the simple system.

The ability to have a universal goal setting methodology has become essential for Cisco Systems, Inc. (Cisco) as it moves towards the implementation of Six Sigma across its manufacturing operations. The company is currently operating under a model that allows each product group, or Business Unit (BU), to define its own FPY goals, but in order to realize the full benefits of the Six Sigma initiative, a coordinated cross-company effort to manage FPY performance is required. As a result an effort to generate a universal FPY goal setting methodology was initiated, starting with the task of creating a methodology to predict FPY for all the systems produced by the company.

Creating the yield prediction methodology needed by Cisco is the central theme of this research work. The hypothesis that is investigated is that the differences in FPY performance for different products can be explained by a number of system and process complexity attributes. Although the analysis is done using exclusively Cisco's manufacturing data, the methodology used to conduct this research and the conclusions generated should be applicable to any electronic systems assembly and test operation.

1.1 Thesis Organization

This thesis is organized into seven chapters as follows:

- Chapter 1 – Introduction: Provides an overview of the thesis, describes the thesis organization, defines the problem statement and the purpose of the study, and enunciates the thesis statement.
- Chapter 2 - Company and Manufacturing Process Overview: Provides background information on Cisco Systems and its manufacturing process.
- Chapter 3 - Research and Literature Review: Presents the literature reviewed as a basis for this analysis.
- Chapter 4 – Research Methodology: Describes the process followed to identify the factors that influence FPY, the creation of a data collection plan and the execution of the data collection process.
- Chapter 5 – Hypothesis Test: Presents the process and results of performing statistical analysis to test the proposed hypothesis.
- Chapter 6 – Conclusions: Provides a summary of the key findings and recommendations
- Chapter 7 – Bibliography: Lists the bibliographical references used for this research

1.2 Problem Statement

Cisco's Manufacturing Operations group requires a universal FPY goal definition methodology to enable the implementation of the Six Sigma program. The Six Sigma initiative is the central component of a company wide effort to facilitate continuous quality improvement through collaboration among product groups, manufacturing, and product design and development. The effort was initiated with the generation of FPY goals for the printed circuit board assembly (PCBA) operation, and needs to be expanded to the second half of the manufacturing process, Direct Fulfillment (DF), the operation where PCBAs are assembled and tested into complete systems.

There are two main reasons why Cisco does not currently have a universal methodology to define system assembly and test FPY goals. First, Cisco has grown not only organically but also through a considerable number of acquisitions, and the manufacturing goal setting methodologies of the acquired companies have not always been fully converted into the existing Cisco practices. Acquired companies typically operate inside Cisco as individual Business Units (BUs), retaining some of their autonomy on designing and executing their manufacturing policies. The process of driving standardization across BUs has recently gained a lot of attention in the company and has been greatly facilitated by the Six Sigma adoption initiative.

Second, Cisco builds a diverse array of networking and communications systems with varying levels of technological and manufacturing complexity. Establishing a universal FPY goal setting methodology faces the challenge of accounting for variations on the expected manufacturing performance due to differences in the products' complexity levels. It is this particular problem that this research work attempts to address.

1.3 Thesis Statement

The main purpose of this thesis is to explore the effect that product and process complexity have on manufacturing performance of electronic systems, in particular those manufactured by Cisco Systems, Inc. To perform this study, the following hypothesis is tested:

Hypothesis: *Variations in System Assembly and Test First Pass Yield (FPY) can be explained by a limited set of product and process attributes, enabling the generation of a yield prediction model that can accurately estimate FPY as a function of them*

If this hypothesis can be confirmed, the resulting yield prediction model can help in the process of defining appropriate FPY performance goals for current and new Cisco products. It would also portray a case study on yield prediction methods that could be used by other companies to understand the impact of product and process complexity on their manufacturing operations.

In the case that the hypothesis is rejected this study would generate insights and recommendations on how to continue to explore the effects of product and process complexity in first pass yield performance.

2 Company and Manufacturing Process Overview

Cisco Systems is the market leader in the networking industry and manufactures more than 4500 different products, each with a wide array of customer selectable options, resulting in a virtually infinite list of potential system configurations. The company's hardware offerings range from relatively simple and inexpensive IP phones to very complex multimillion-dollar network routing systems. Manufacturing operations span the globe and are almost 100% outsourced. A high level overview of the typical manufacturing process is described in the next section.

2.1 Cisco's Manufacturing Process

Cisco manufacturing process is summarized in Figure 1. At a very high level there are two main steps in the process: Printed Circuit Board Assembly (PCBA) and Direct Fulfillment (DF). During the PCBA operation, electrical components are positioned and soldered in place on a dielectric substrate that contains the conductive tracks to form an electrical circuitry (Printed Circuit Board or PCB). Once all the components are secured in place, the PCBA goes through a testing process that is usually divided in two main steps, in-circuit test and functional test. In-circuit test verifies the structural integrity of the board, meaning that it checks for problems with the electrical connections between the board and the components. Common defects captured by this test are solder joint opens and shorts. Functional test, on the other hand, verifies the correct functionality of the board by exercising its components via diagnostic scripts. During this test, the board is tested at corner voltage and temperature conditions to ensure that the board remains fully functional according to its specifications. Once a board has successfully passed all tests it is sent to continue processing at a Direct Fulfillment location.

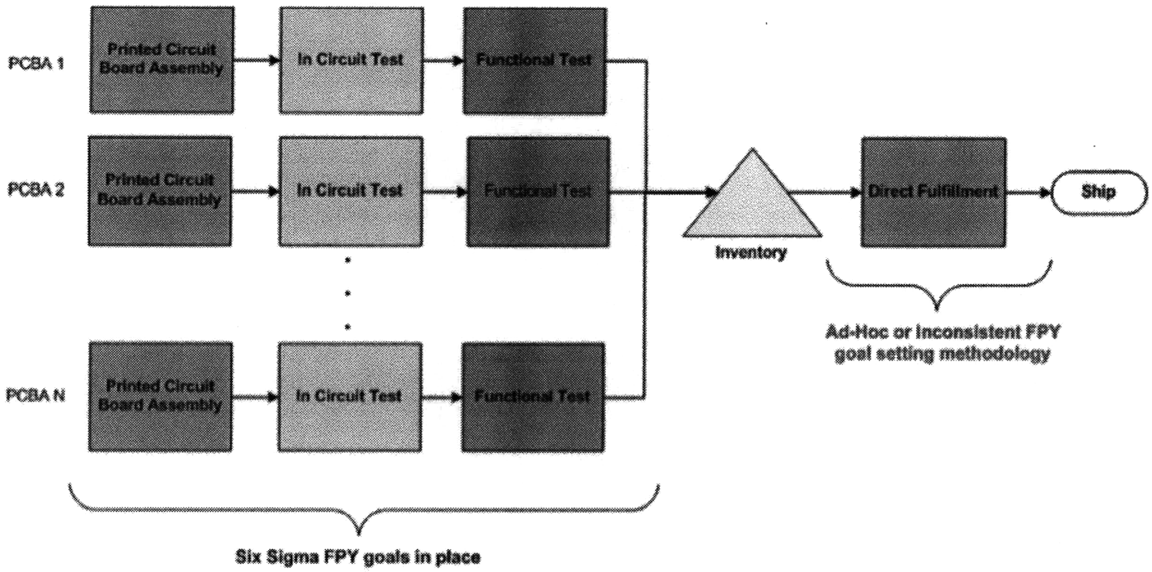

PCBA 1
PCBA 2
PCBA N
Printed Circuit Board Assembly
In Circuit Test
Functional Test
Inventory
Direct Fulfillment
Ship
Ad-Hoc or Inconsistent FPY goal setting methodology
Six Sigma FPY goals in place

Figure 1. High Level Manufacturing Flow

Direct Fulfillment (DF) is the second and final component of the manufacturing process. PCBAs arrive at the DF location to be either assembled into a configured to order system or shipped as a stand-alone product (spare). Systems are usually composed of a chassis, one or more PCBAs, power supplies and other electronic components such as hard drives, memory modules etc. Once assembled, the systems undergo a series of tests that verify that the equipment has been assembled correctly and that it is functioning as expected. The assembly test verifies the correct configuration of the system and the quality of the connections between the different boards and components. The system functional tests verify that the system is functioning as specified and usually exercises portions of the PCBA circuitry that cannot be tested at PCBA level tests. After each test operation, in the PCBA and DF process, if a PCBA or system fails it is sent to a debug and rework flow in which the failure is analyzed and necessary repairs are performed.

Once the system has successfully passed all test operations it is shipped to the customer. When a PCBA is to be shipped as a spare it is usually assembled into a golden system and processed through the regular testing flow. When the golden system successfully passes all tests, the PCBA is removed and shipped to the customer.

Different yield metrics, such as Rolled Throughput Yield (RTY) and First Pass Yield (FPY) are continuously reviewed in order to understand the performance of the manufacturing operations. Figure 2 gives an example on how this metrics are calculated. RTY is the probability that a unit can pass through the whole manufacturing process without any defects and is calculated by multiplying the yield for each one of the test operations. FPY is the ratio between the number of systems that passed all test operations without needing any debug or rework and the total number of systems that went into the process. This study focuses only on FPY to maintain a consistent approach with the previous Six Sigma efforts in the company.

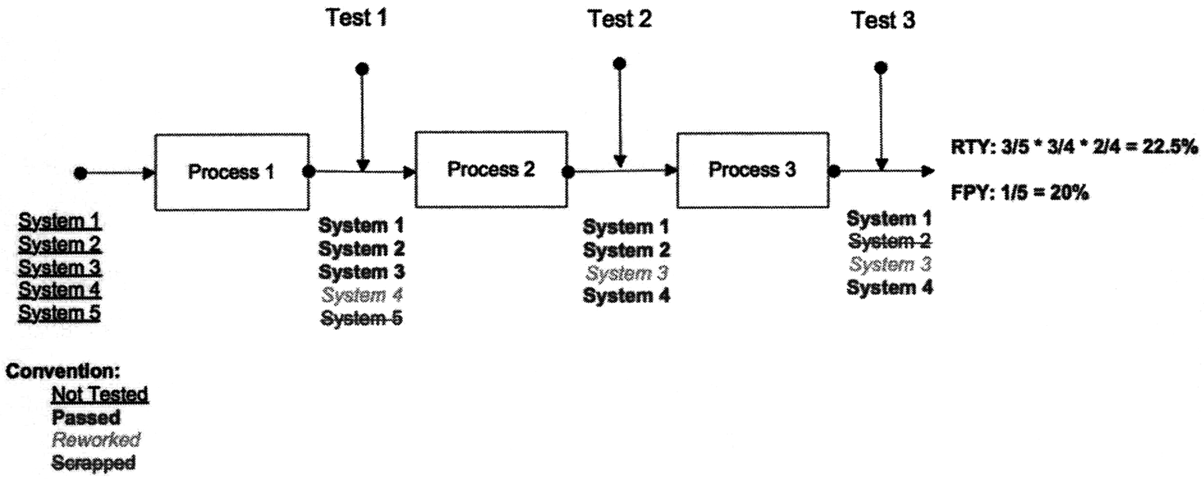


Figure 2. RTY and FPY Calculation Example

3 Research and Literature Review

Yield estimation in the electronics industry has been the subject of multiple studies in the industry and academia due to its importance in managing manufacturing cost, line capacity, materials procurement and on-time delivery. There are in particular many publications on yield prediction for Integrated Circuits manufacturing (ICs) and Printed circuit Board Assemblies (PCBA) manufacturing. There is, however, not a significant number of references on estimating electronic systems assembly and test yield.

IC manufacturing yield prediction has been studied thoroughly and continues to be a developing field. Although many methods have been explored, the common theme across them is the creation of mathematical models (usually statistically based) that predict yield performance based on the defect density of the design on a given process and the fault coverage (percentage of defects detected) of the test suite used during manufacturing. Different approaches have been proposed on how to effectively estimate those two parameters, including historic process data, circuit design and layout characteristics, and test coverage measurements at different stages of the design process. Despite the complexity of today's IC designs, manufacturing processes yield estimation has benefited from the existence of substantial amount of process data to estimate defect densities and the increasing amount of simulation power to measure test coverage.

In general, two different methods have been explored to predict PCBA manufacturing yield: process yield estimation, and board design yield estimation (F. Helo, 2000). Process yield estimation uses a very similar approach to the one described above for IC yield prediction, where defect density and fault coverage are the main factors used to build a statistical model. In this case, defect density is approximated by the defect rate of the different components (percentage of faulty components shipped by suppliers), and fault coverage is calculated as the percentage of defects that can be detected by the board tests. The efficiency of this method has been hindered by two main factors:

The need to rely on suppliers to accurately report the defect rates of the components they provide, and the difficulty of accurately measuring the fault coverage for board tests.

PCBA design yield estimation proposes models that incorporate PCBA design attributes such as number of components, board area, number of solder joins, and board circuit layer count to generate statistical yield prediction models. It has been demonstrated (Y. Li, 1994) that as board design complexity increases, the expected manufacturing yield for the board decreases. Li et al studied the effect of board design parameters on PCBA yield by relating the design attributes and yield performance of 30 different PCBAs through regression analysis and artificial neural networks. The resulting models were able to predict yield with a root mean square (RMS) error of less than 5%.

Cisco performed a similar statistical analysis using manufacturing data for its own products finding similar results (Averbeck, 2008). These results have enabled the generation of a PCBA yield goals generation methodology that uses the yield estimation model to translate different board attributes into Six Sigma defect opportunities. The definition of what constitutes a defect opportunity is a fundamental step in defining Six Sigma goals of 3.4 defects per million opportunities (DPMO) (Shina, 2002).

No public references were found on methods to estimate system assembly and test yield. The common practice in the industry seems to be to estimate expected yield performance based on historical data and the past performance of systems with similar components, architecture and manufacturing processes. The fact that this approach has been used with some success in the industry appears to support the idea that there is a set of system complexity attributes that have a direct influence on how the systems perform in the manufacturing line. The existence of this relationship, between product and process attributes and yield performance, is precisely the thesis that this study explores.

To explore the validity of this thesis an approach similar the PCBA design yield estimation was used. The study starts by identifying the design and process attributes that may have an impact on yield and later investigates the possibility of relating them to the system's actual yield performance by using statistical tools such as regression analysis and artificial neural networks.

4 Research Methodology

This chapter describes the process followed to identify the main product and process attributes that influence FPY, and reviews the creation and execution of the data collection plan. The data collection plan generated the information used to perform the statistical tests required to identify the relationship between the selected factors and FPY.

4.1 Identifying the Factors that Influence First Pass Yield

The identification of the different factors that influence system assembly and test FPY was the first step required to start with the study. Managers and engineers agreed that product and process complexity had a direct impact on the yield performance observed at the different factories. However, the definition of what constituted complexity at the product and process level was not clear. When asked to compare various products, managers and engineers seemed to have no issues in ranking them using subjective measures of complexity. Yet, when asked why a particular system ranked higher than other the answers where not always clear and consistent: it was difficult to describe what made a product more complicated.

It was evident then that it was required to perform an exercise to determine as precisely and comprehensively as possible the product and process complexity factors that have a direct impact on a system's manufacturing yield. The result should be a list of measurable attributes that capture the subjective reasoning used by the different individuals to assess manufacturing complexity. To facilitate the process of generating this list a Cause-and-Effect (also known as Fishbone or Ishikawa) diagram was created through a brainstorming and prioritization exercise.

The resulting diagram in Figure 3 reveals six mayor categories of factors that influence FPY performance for Cisco products: Assembly and Test Process, Product Complexity, Product Design, Production Volume, and Incoming Material Quality. Each category has one or more individual factors identified by the subject matter experts to have an impact on the manufacturing yield performance of a system.

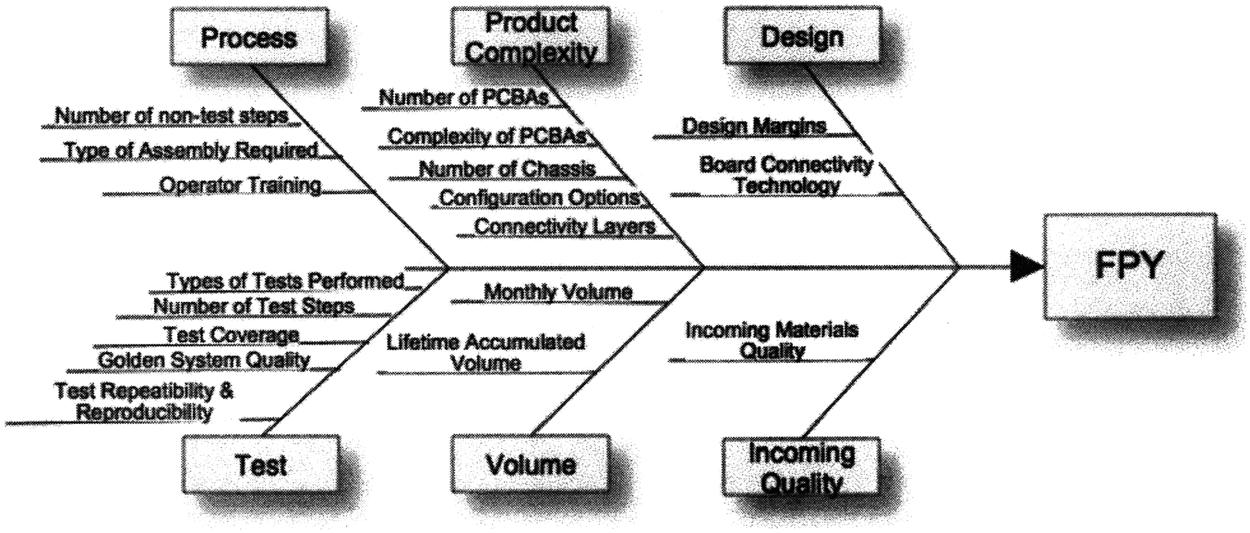
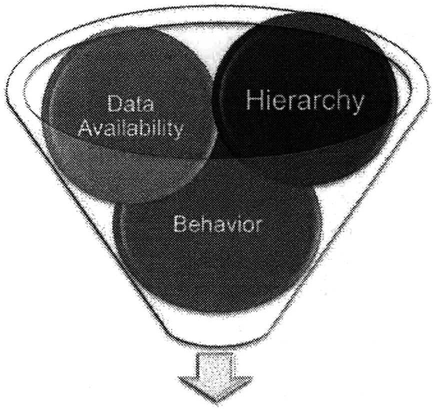


Figure 3. Cause and Effect Diagram for System Assembly and Test FPY

Further analysis was required to select the yield drivers that should be used to test the hypothesis. Not all of the factors identified through the exercise fell into the scope of this project, which attempts to deal with process and product complexity factors. There are also drivers that are either hard to measure or for which data is not available or difficult to obtain. To identify the factors relevant to this study, a framework based on three main selection criteria: factor hierarchy, data availability and organizational behavior was used. The details of the framework and the resulting factor list are covered in detail in the next section.

4.2 Creating a Data Collection Plan

Extracting the relevant factors from the general yield drivers list required a formal framework to guarantee that no significant factor was left out and that all factors selected were significant to the objectives of the study. The framework used (Figure 4) consists of three main criteria: factor hierarchy, data availability and organizational behavior.



Inputs to data collection plan

Figure 4. Factor Selection Criteria

By factor hierarchy it is understood that there are factors that are a consequence or a function of other higher hierarchy factors. By looking at the hierarchical relationship between the factors it is possible to intend to identify and select those factors that are at the top of the scale. By using only the principal factors, those that are the cause of product and process complexity and not the result, it is possible to better focus the analysis, minimizing redundancy and correlation among variables.

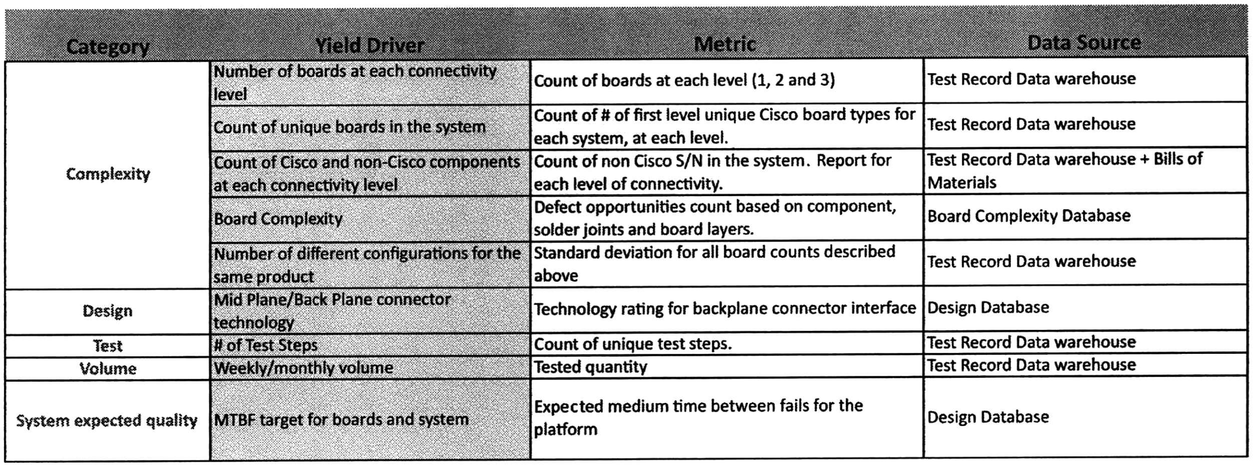
Number of assembly steps is an example of a factor that was removed from the list of drivers due to this criterion. Although there was consensus on its impact on yield performance, it was determined by the team that the number of assembly steps should be a function of the number of boards at each connectivity level in the system. In this case, number of boards at each level had a higher hierarchy than assembly steps and, a result, number of boards at each level was included in the final list but number of assembly steps was not.

The second criterion used was data availability and data collection feasibility. A metric was defined for each factor affecting FPY and the possibility of constructing it with the available data was assessed. Those factors that could not be measured using the existing data sources where excluded from the analysis. Two relevant factors that could not be included in the analysis because of challenges collecting the necessary data were test coverage (percentage of defects that could be successfully screened out by the test process) and incoming component quality (defects per million, or DPM, for different components in the system).

The third criterion used to evaluate the different factors was their relationship to organizational behaviors. If a given factor was the result of poor organizational behavior it was removed from the analysis, helping maintain the study focused on product and process complexity. Examples of factors that fell into this category are: Tool wear and tear (poor preventive maintenance policies), design marginalities (inadequate design procedures and controls) and quality of operator training.

Another reason to exclude these factors from the study is that the ultimate goal of this effort is to facilitate the creation of an FPY goal setting methodology. To serve its purpose of driving quality improvements across the organization, the goal generation methodology should not include any yield loss allowances for poor organizational performance. Removing all behavioral related factors from the FPY estimation model prevents those factors from being included in the goal generation methodology.

The results of applying the three principles described above were captured in a Data Collection Plan (Figure 5). This document summarizes the selected factors and the metrics used to quantify them and constitutes the starting point for the data collection process that is described in the following section.



Category	Yield Driver	Metric	Data Source
Complexity	Number of boards at each connectivity level	Count of boards at each level (1, 2 and 3)	Test Record Data warehouse
	Count of unique boards in the system	Count of # of first level unique Cisco board types for each system, at each level.	Test Record Data warehouse
	Count of Cisco and non-Cisco components at each connectivity level	Count of non Cisco S/N in the system. Report for each level of connectivity.	Test Record Data warehouse + Bills of Materials
	Board Complexity	Defect opportunities count based on component, solder joints and board layers.	Board Complexity Database
	Number of different configurations for the same product	Standard deviation for all board counts described above	Test Record Data warehouse
Design	Mid Plane/Back Plane connector technology	Technology rating for backplane connector interface	Design Database
Test	# of Test Steps	Count of unique test steps.	Test Record Data warehouse
Volume	Weekly/monthly volume	Tested quantity	Test Record Data warehouse
System expected quality	MTBF target for boards and system	Expected medium time between fails for the platform	Design Database

Figure 5. Data Collection Plan

4.3 Data Collection

The Data Collection Plan was executed in order to gather the data necessary to perform the study. Data collection was performed for all Cisco systems being manufactured by the company's contract manufacturers in the period between July 27 and October 25 2008 (4th quarter of Cisco's 2008 fiscal year). The decision to collect data for all systems was motivated by the risk of biasing the sample due to the lack of previous experience selecting products based on the selected attributes.

The total number of systems to be analyzed rounded the nine million count, requiring the development of an automation tool to make the process feasible and avoid introducing errors due to manual data collection and reporting. The resulting automation tool also enables experiment repeatability, allowing subsequent studies of production data to be performed.

The raw data produced by the automation tool was then manually scrubbed to remove records that could affect the effectiveness of the analysis. In particular, records with less than 30 samples or with incomplete information were removed. After eliminating problematic entries the resulting data set consisted of 232 records. This data set was used to perform the statistical analysis necessary to understand how well the factors identified explained variations in yield performance for the selected systems. The statistical analysis is covered in detail in the next section.

5 Hypothesis Test

This chapter presents in detail the statistical analysis performed to understand the relationship between the system and process complexity attributes identified in the previous chapter and the systems' FPY performance. First, basic first and second order linear regression models are explored for their fit and prediction capabilities. Subsequently, more elaborate non-linear models, logistic regression and artificial neural networks, are analyzed.

5.1 Statistical Analysis

Statistical analysis was the main tool used to test the ability to predict FPY performance based on process and product complexity attributes. Two particular factors were taken into account in deciding if enough evidence existed to prove or disprove the hypothesis: The regression's R-Squared value and the Cross Validation's RMS (Root Mean Square) error.

The Coefficient of Determination, or R-Squared, is a common measure of goodness of fit for regression analysis. In this case, it indicates the percentage of the variation on the FPY performance that can be explained by the complexity attributes used in the regression model. Based on previous experiences at Cisco and published literature on statistical based yield prediction models, values above 0.7 or 0.8 are acceptable levels of R-squared (Averbeck, 2008)(Y. Li, 1994). Regression results were compared to this generally accepted level when deciding if the hypothesis could be confirmed or not.

Root Mean Square Error (RMSE) is used to measure the predictive accuracy of the model. RMSE was chosen to facilitate the comparison of the results with the different yield prediction models found in the literature. To calculate the RMSE, the model resulting from the statistical analysis is used to predict the FPY of a number of data points not used during the model generation process. Then, the RMS difference between the predicted and actual yield is calculated. The smaller the RMS error, the better is the model for prediction purposes. Literature (Y. Li, 1994) (F. Helo, 2000) reports RMS errors of 3% to 5% in successful attempts to create yield prediction models, hence similar levels were considered to constitute acceptable evidence to support this study's hypothesis. The formula used to calculate RMSE is described in Equation 1.

$$RMS = \left(\sum_{i=1}^{N} (Y_i - Y_i^p)^2 / N \right)^{0.5}$$

Y_i is the actual yield and $Y_i P$ is the predicted yield

Equation 1. RMS Error

To create and test the model, the data points were divided in two sets, a fitting set and a testing set. The fitting set was used to run the regression and build the model, and the test set was used to test the model for prediction accuracy. The selection of the data points that go in each model was done randomly to avoid sample bias.

The following sections describe in detail the model generation and validation process for each one of the statistical modeling techniques explored. The last section of this chapter summarizes and analyzes the findings of the analysis.

5.1.1 First Order Linear Regression

The first order linear regression model described in Equation 2 was explored first. This model represents the dependent variable (Y), in this case FPY, as a linear combination of first order independent variables (x_1, x_2 .. x_k), in this case the complexity attributes. The model includes also an additional term, ε, a random variable that accounts for everything else that influences Y but is not captured by $x_1 \dots x_k$.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Equation 2. First Order Linear Regression Model

Three main steps were followed when creating the model: 1) Factor selection, 2) Regression execution, and 3) Results analysis and cross validation.

Factor selection is used to identify the simplest model that achieves the best possible fit. Reducing the number of factors in the model also helps prevent instability in the equation's coefficients due to multicollinearity. In this analysis, factor selection was performed using two different tools: correlation matrix, and all subsets regression analysis. The correlation matrix analysis identified the factors with strong correlation with other factors in the model. Factors that exhibited strong correlation were removed from the pool of variables that was later used to execute the regression analysis. Afterwards, during the regression execution phase, the Variance Inflation Factors (VIF) for the each of the independent variables in the model was calculated to confirm that the model indeed had no multicollinearity problems. VIF values of 5 or less were considered acceptable.

The reduced factor group was then used to perform an all subsets regression analysis. The all subsets regression analysis helps to identify the best fit attainable with all the possible subsets of the factors in the model. The analysis was performed using all the available data points (232), A regression model was generated for each subset and R-Squared and Adjusted R-Squared were computed for each model. The factor subset that generated the model with the highest Adjusted R-Squared was selected to continue with the analysis. In this case, Adjusted R-Squared was used instead of R-Squared because it adjusts the goodness of fit calculation for to the differences in factor count on each model (Patel, 2003).

The resulting model contains nine factors total, down from 25 first order variables available in the data set. Ten factors were removed due to multicollinearity, five factors dropped as a result of the all subsets regression analysis (Table 1), and one additional factor was later removed because it was not significant according to the regression ANOVA results (p-value > 0.05). The specific factors remaining in the model and the model itself are not included in this document for proprietary reasons.

The regression results for the first order linear regression are summarized in Figure 6. R-Squared for the model is 65%, slightly below the 70 % target established as acceptable to confirm our hypothesis in terms of model fit.

Vars	R-Sq	R-Sq (adj)
1	46.3	46.1
2	53.4	52.9
3	55	54.5
4	57.2	56.4
5	59.1	58.2
6	60.4	59.3
7	62.4	61.2
8	64.5	63.2
9	65.4	64
10	65.7	64.2
11	65.9	64.2
12	66.1	64.3
13	66.3	64.3
14	66.3	64.2
15	66.3	64

Table 1. Best Subset Analysis for 1st Order Linear Regression

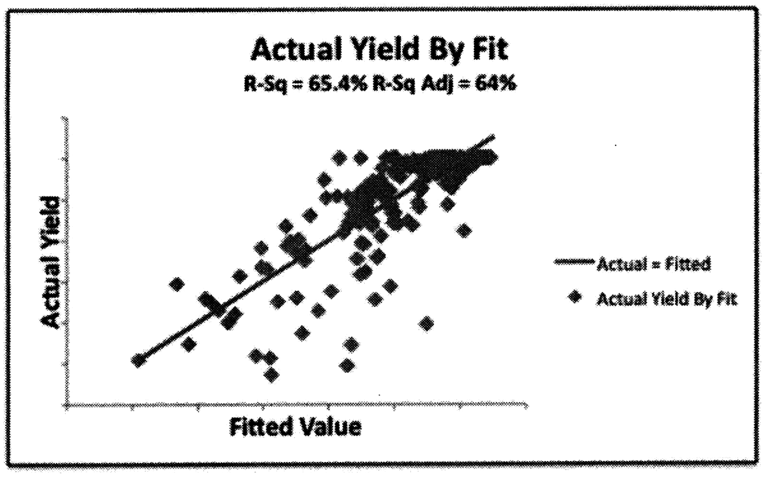


Figure 6. 1st Order Linear Regression Results Summary

5.1.1.1 Regression Cross Validation

The model was tested for its predictive capabilities by dividing the 232 data points in two sets, a training set to generate the fit equation and a testing set for testing it. The training set contains 174 points and the testing set contains the remaining 58. The points in the testing set were randomly selected, but a consideration was made to guarantee that the testing set had no points outside the range of the training set.

As mentioned before, RMS error was used to measure the model's predictive accuracy. Three cross-validations were performed by rotating the training and testing sets. For each cross-validation, the RMS error was calculated using Equation 1, and the model's overall RMS error was calculated as the average for the three rotations.

Table 2 and Figure 7 summarize the results of the model cross-validation. When used for predictive purposes the first order linear model exhibits degraded performance, with R-Squared shrinking by 9% on average. The average RMS error is 8.51%, not meeting the 5% goal defined as acceptable during the hypothesis formulation.

	Fit Model R-Sq	**Prediction R-Sq**	**Prediction RMSE**
Rotation 1	64%	65%	8.28%
Rotation 2	69%	53%	9.26%
Rotation 3	68%	56%	8.00%
Average	**67%**	**58%**	**8.51%**

Table 2. 1st Order Linear Regression Cross Validation Results

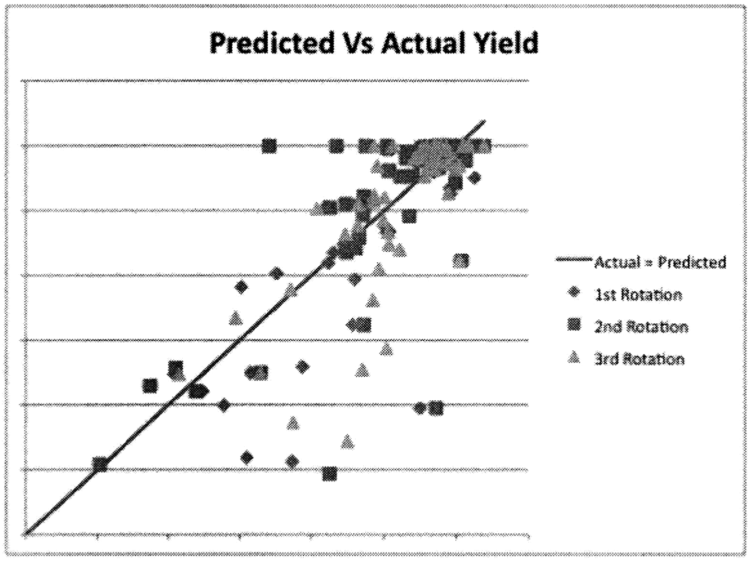


Figure 7. First Order Linear Regression Prediction Performance

5.1.2 Second Order Linear Regression

The second order linear regression model was explored next. This model is similar to the first order model already analyzed in the previous section, with the difference that, in addition to the first order factors, it includes second order and interaction terms for the independent variables. Equation 3 shows the general form of the second order regression model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_{11} x_1^2 + \ldots + \beta_{kk} x_k^2 + \beta_{12} x_1 x_2 + \ldots + \beta_{(k-1)k} x_{k-1} x_k + \varepsilon$$

Equation 3. Second Order Linear Regression Model

The process of building and validating the model was identical to the process followed for the first order model. Factors were screened out via correlation matrix and best subset analysis, selecting the combination that produced the highest Adjusted R-Squared value. The model was tested using R-square as a measurement of goodness of fit and RMS error as a measurement of prediction accuracy.

The model generated contains a total of 13 factors, four of which are first-order, three are second-order and the remaining six are factor interactions. 16 factors were selected as a result of the best subset analysis (Table 3) and 3 factors were later removed as the regression ANOVA identified them as not significant (p-value > 0.05). Again, as with the first order model, factor details and model coefficients cannot be included in this document for company confidentiality reasons.

The regression results for the second order linear regression are summarized in Figure 8. The model represents an improvement over the simpler first order model, with R-Squared of 71.6% Vs 65%. The model passes the test criteria of 70% R-Squared established as acceptable to confirm the hypothesis in terms of goodness of fit.

Vars	R-Sq	R-Sq (adj)
1	46.3	46.1
2	54.5	54.1
3	59.1	58.6
4	61.2	60.5
5	62.8	62
6	65.3	64.4
7	66.2	65.1
8	67.3	66.1
9	68.4	67.1
10	69.5	68.1
11	70.3	68.8
12	71.1	69.5
13	71.6	69.9
14	72	70.2
15	72.4	70.5
16	72.8	70.8
17	73	70.9
18	73.2	70.9
19	73.3	71
20	73.4	70.9
21	73.5	70.9
22	73.5	70.8
23	73.6	70.6
24	73.6	70.5
25	73.6	70.4
26	73.6	70.2

Table 3. Best Subset Analysis for 2nd Order Linear Regression

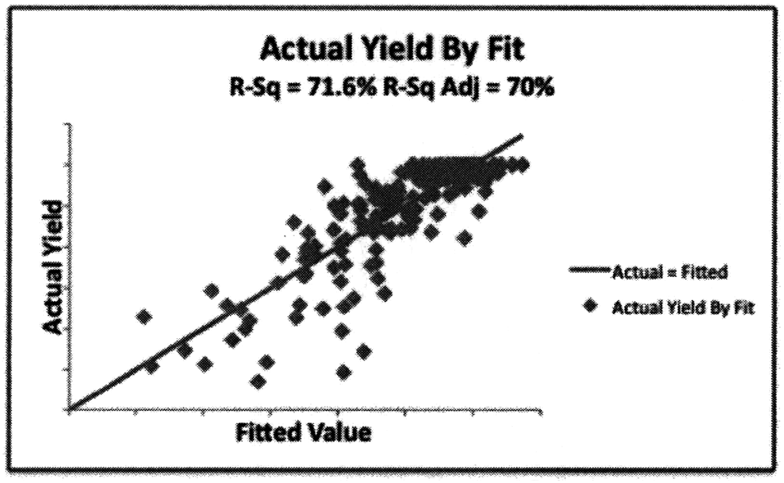


Figure 8. 2nd Order Linear Regression Results Summary

5.1.2.1 Regression Cross Validation

To assess the general capabilities of the model in terms of its prediction accuracy the same process as the one followed for the first order regression was used. The model's prediction accuracy was measured by calculating the RMS error for three different cross-validations using rotations of the training and testing data sets. The model's overall RMS error was calculated as the average for the three rotations.

Table 4 and Figure 9 summarize the results of the model cross-validation. When used for predictive purposes, the resulting model exhibits better performance than the first order model. Fit (R-Squared) shrinks 6% on average compared to 9% for the first model. RMS error of the prediction also improves, going from 8.51% to 7.64%. However, the average RMS error is still higher than the 5% level defined as acceptable during the hypothesis formulation.

	Fit Model R-Sq	**Prediction R-Sq**	**Prediction RMSE**
Rotation 1	70%	75%	7.34%
Rotation 2	74%	62%	8.25%
Rotation 3	74%	63%	7.33%
Average	**73%**	**67%**	**7.64%**

Table 4. Second Order Linear Regression Cross Validation Results

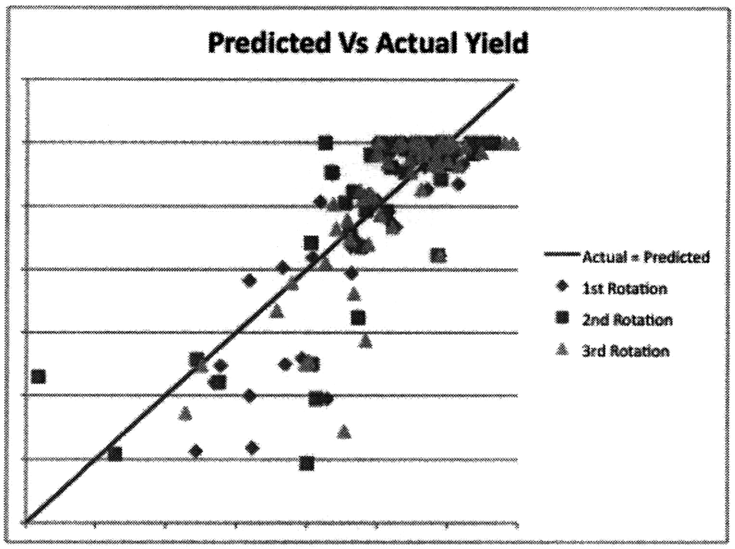


Figure 9. Second Order Linear Regression Prediction Performance

5.1.3 Logistic Regression

Logistic regression is an extension of the traditional linear regression techniques and is particularly useful when the dependent variable is binary. This regression methodology is appropriate for this study because yield can be interpreted as the probability of a device passing (or failing) the different tests steps along the manufacturing process. Traditional multiple linear regression may be failing to produce better results because it generates predictions that fall outside the valid probability range 0 to 1. The general form for the logistic regression is described in Equation 4. This model is extensively used in econometrics and in life sciences, particularly in modeling risk factors in epidemiology (Patel, 2003).

$$\text{Prob}(Y = 1 \mid x_1, x_2 \cdots x_k) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

Equation 4. Logistic Regression Model

The same first, second order and interaction factors used in the linear regression models were used in constructing the logistic model. The model contains 13 factors total (4 first-order, 3 second-order and 6 are interactions). The regression results for the logistic regression are summarized in Figure 10. The model represents a slight improvement over the linear regressions, with R-Squared[1] of 73,6%. As with the second order linear regression, the model passes the fit criteria of 70% established as acceptable to confirm our hypothesis in terms of goodness of fit.

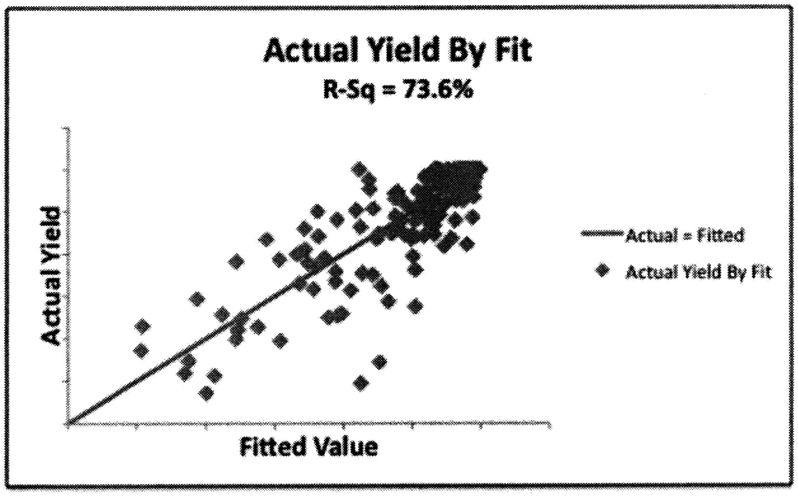


Figure 10. Logistic Regression Results Summary

5.1.3.1 Regression Cross Validation

The same cross validation methodology used for the linear models was used to validate the general prediction capabilities of the model generated through logistic regression. Three different random training and validation data sets were used to measure the prediction RMS error. The overall RMS prediction error for the model was calculated as the average of the three cross-validations.

[1] Goodness of fit for the logistic regression is usually measured using the deviance of the model. However, in this case, to facilitate the comparison of the logistic model with the previous models, R-Squared was calculated. R-Squared was calculated as the square of the correlation between the actual yield values and the fitted values from the model.

Table 5 and Figure 11 summarize the results of the model cross-validation. When used for predictive purposes the resulting model exhibited in general similar characteristics to the second order linear model. Fit (R-Squared) shrinks on average 7% and RMS error is 7.8%. Both metrics represent a slight degradation in performance when compared to the second order model, which presented 6% R-Squared shrinkage and 7.6% RMS error. As noted when analyzing the results for the linear models, the average RMS error is still higher than the 5% level defined as acceptable during the hypothesis formulation.

	Fit Model R-Sq	Prediction R-Sq	Prediction RMSE
Rotation 1	71%	72%	7.50%
Rotation 2	76%	66%	8.50%
Rotation 3	73%	61%	7.54%
Average	**73%**	**66%**	**7.85%**

Table 5. Logistic Regression Cross Validation Results

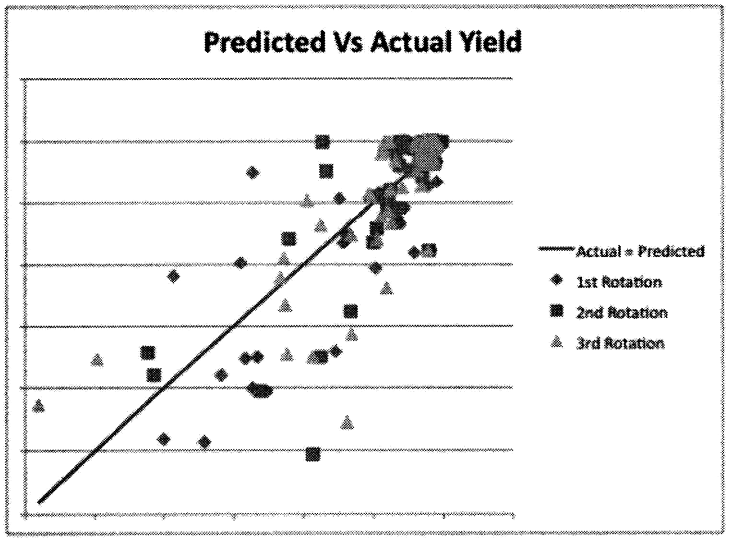


Figure 11. Logistic Regression Prediction Performance

5.1.4 Artificial Neural Network

Artificial Neural Network (ANN) modeling was the last technique explored. An artificial neural network is a mathematical simplification of the structure of a biological neural network. The model has been successfully used in various areas, including machine learning, financial markets performance prediction, and other complex data mining applications. The main advantages of neural networks are their ability to effectively model almost any type of response surface and their generalization capability. Generalization means that the resulting models can usually perform very well when used with data points that do not resemble closely the data set used to create, or "train", the model. Because the linear and logistic models failed to meet the predefined prediction error goals, leveraging on the ANNs generalization capabilities seemed the next logic step in exploring the hypothesis.

The ANN model that was used has the form described in Figure 12. The structure consists of three layers: one input layer, one hidden layer and one output layer. Each layer is composed by a number of nodes. The input layer has as many nodes as independent factors are in the model, and the output layer has as many nodes as dependent variables. In this case the model uses the same 13 factors used in the second order and logistic models, generating 13 nodes for the input layer. Since there is only one dependent variable, FPY, the output layer has only one node.

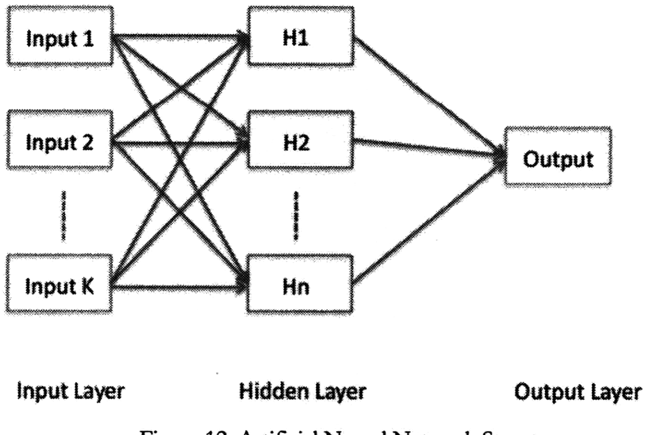


Figure 12. Artificial Neural Network Structure

Each node in the hidden layer contains a nonlinear function of the input factors. In this case the nonlinear function is the logistic function, also known as the sigmoid function. This is the same function used in the logistic regression (Equation 5). In general, each one of the hidden nodes H_j is defined as indicated in Equation 6, and the coefficients β and the constant c_j are estimated during the training process. There is no clear theory to guide the selection of the number of nodes in the hidden layer, however, a network with too few hidden nodes may not be able to learn from the data, and one with too many nodes may overfit (memorize) the training data and loose its generalization capabilities. The common practice is to use trial and error to reach the configuration that performs as desired. In this case, the optimal number of nodes was found to be two.

$$S(x) = \frac{1}{1 + e^{-x}}$$

Equation 5. Sigmoid Function

$$H_j = S_H\left(c_j + \sum_{i=1}^{k}\left(\beta_{ij} X_i\right)\right)$$

Where k is the number of independent variables and S_H is the sigmoid function

Equation 6. Hidden Node Equation

The output node contains a function that aggregates the results of each one of the hidden nodes and generates the response variable, in this case the predicted FPY. The output node function is described in Equation 7. The coefficients α and the constant d are estimated during the training process.

$$Y = S\left(d + \sum_{j=1}^{n}\left(\alpha_j H_j\right)\right)$$

Where n is the number of hidden nodes and S is the sigmoid function

Equation 7. Output Node Equation

Cross-validation was performed in parallel to the training process of the neural network; hence the results for both the training and cross-validation processes are presented consolidated in the following graphs and tables. Cross-validation was performed by randomly holding back 25% of the available data points, generating training and testing sets equivalent to the ones used in the linear and logistic regressions (174 points in the training set and 58 in the cross validation set). Again, three different cross-validations were performed.

Figure 13 shows the fit performance of the model and Table 6 summarizes the cross validation results. The neural network has superior performance compared to the linear and logistic regressions, exceeding the criteria defined for both fit and prediction accuracy. R-squared for the resulting model is 85% and the average RMS prediction error is 4.8%. R-squared shrinkage is significant (12%) when comparing the training and cross-validation results, the highest for all the models explored. However, despite the significant shrinkage, the ANN is the only model studied in which both the training and validation R-Squared values are higher than 70%.

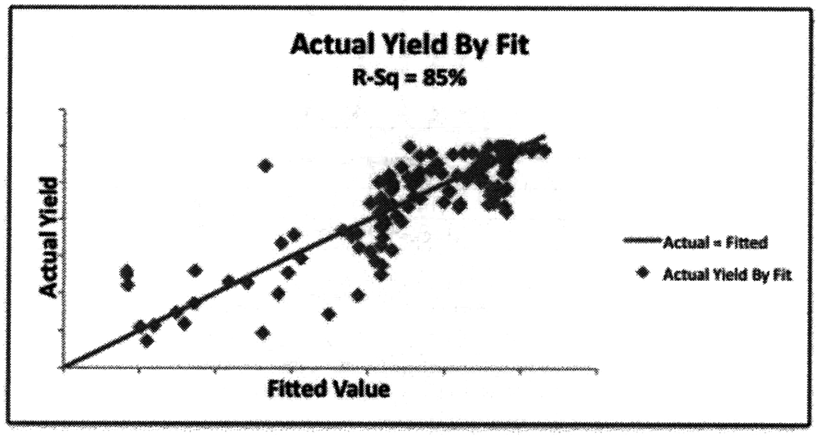


Figure 13. Neural Network Performance Summary

	Fit Model R-Sq	Prediction R-Sq	Prediction RMSE
Rotation 1	85%	73%	4.79%
Rotation 2	85%	72%	4.80%
Rotation 3	85%	75%	4.80%
Average	**85%**	**73%**	**4.80%**

Table 6. Neural Network Cross Validation Results

5.2 Results Analysis and Summary

The statistical modeling performed has shown that it is possible to predict FPY performance based on selected product and process complexity attributes. The analysis confirms the hypothesis outlined at the beginning of this study: "*Variations in System Assembly and Test First Pass Yield (FPY) can be explained by a limited set of product and process attributes, enabling the generation of a yield prediction model that can accurately estimate FPY as a function of them*"

Table 7 summarizes the key results for the statistical analysis performed. According to the criteria defined (R-Squared > 70% and RMSE < 5%), satisfactory prediction levels were achieved only by the utilization of neural network modeling techniques, as linear first and second order, and logistic regression models were not able to conform to the specified 5% maximum RMS prediction error.

Model	Fit R-Sq	Prediction R-Sq	Prediction RMSE
Linear 1st Order	65%	58%	8.51%
Linear 2nd Order	72%	67%	7.64%
Logistic	74%	66%	7.85%
Neural Network	85%	73%	4.80%

Table 7. Summary of Statistical Analysis Results

In general, both the goodness of fit and the generalization capabilities improved as the model complexity increased, with the exception of the change from a second order linear model to the logistic model, where the prediction RMS error degraded slightly. In this case, even though the increased complexity did not result in improved prediction capabilities, the logistic model is an improvement over the linear model as it guarantees that the resulting yield prediction is on the 0 to 1 range.

The fact that the neural network outperformed the other modeling techniques is not surprising. In addition to the ability to model non-linear relationships effectively, the flexibility of the neural network allows for subtle relationships between the independent variables and the response variable to be effectively captured. The neural network model is also good at dealing with the noise present in the data set, as there are a number of variables not related to product complexity, such as test record completeness, differences in manufacturing partners operating methods, and process capability factors that were not considered in this analysis.

6 Conclusions and Recommendations

This study has demonstrated that differences in system assembly and test FPY performance can be explained as a function of product and process complexity attributes. Besides confirming the initial hypothesis outlined at the beginning of this study, this result is important because it provides the foundation to create FPY goal definition methodologies that take into account product complexity attributes. Defining complexity based yield goals enables cross product benchmarking, identification of best practices, and the execution of yield improvement initiatives.

Along with the confirmation of the hypothesis, there are other important observations and recommendations that need to be mentioned. The generation of an acceptable FPY prediction model required the utilization of complex, non-linear data mining techniques. Of the four methods explored (first and second order linear regression, logistic regression and artificial neural networks), only artificial neural networks generated a yield prediction model that met the initial thresholds for acceptable goodness of fit and prediction accuracy. The use of neural networks makes the process of understanding the impact that each factor has on the dependent variable a very complex one because there is an intermediate non-linear layer, rather than a direct path, from the input variables to the output variable.

Not having a good understanding of the relationship between each one of the complexity factors and FPY may limit the effectiveness of the goal definition methodology. One of the main objectives of relating product and process attributes to FPY is to identify leverage points in the product design and manufacturing process that can help improve manufacturing performance. However, the ability to drive manufacturing friendly designs and processes depends on the possibility of explaining how process and design decisions affect product performance. Generating a good level of understanding of the impact that each one of the factors in the neural network has in the model output is not an easy task and would require significant effort from the company. Some techniques that may be used to accomplish this task are: varying the levels of the different inputs to the network, and the usage of trees.

From this perspective, it may be worth to invest time and resources on investigating the possibility of improving the yield prediction capabilities of the linear models explored in this study. These models could be improved by focusing on understanding and controlling factors not related to product complexity that may be interfering with the model accuracy and the FPY performance of the products. Examples of such factors include test record completeness and accuracy, test defect coverage, incoming quality levels, variations in manufacturing procedures across locations, and out of control processes among others. If, after careful analysis, the effectiveness of the simpler models cannot be improved, management should weight the option of sacrificing model accuracy for the possibility of facilitating the task of driving the desired organizational performance.

Finally, although this study was performed using Cisco Systems' product and manufacturing data, the general process outlined in this exercise should be applicable to solve similar problems in other companies and industries. The core components of the methodology presented can be easily reproduced: 1) identify the key complexity attributes, 2) design and execute a data collection plan and 3) generate statistical models to test the validity and impact of the selected factors.

BIBLIOGRAPHY

7 Bibliography

Ambler, D. F. (1997, July-September). The Economics of System-Level Testing. *IEEE Design & Test of Computers* , pp. 51-58.

Averbeck, L. (2008). Manufacturing Yield Management. *Internal Cisco Systems Document* . Manufacturing Operations, Cisco Systems.

Byle, F. (2001). Using Industry DPMO Standards - An In-Depth Look at IPC-9261 and IPC-7912. *Proceedings of the technical progtam: SMTA International* (pp. 507-510). Chicago: SMTA.

Chen, M. M. (1994). Defects, Fault Coverage, Yield and Cost, in Board Manufacturing. *International Test Conference* (pp. 539-547). IEEE.

D. Ciplickas, S. F. (2001, October). A New Paradigm for Evaluating IC Yield Loss. *Solid State Technology* .

Dooley, B. J. (1983, January). A Model for the Prediction of Assembly, Rework, and Test Yields. *IBM Journal of Research and Development* , 59-67.

E. Kamen, A. G. (1999). Analysis of Factors that Affect Yield in SMT Assembly. *National Electronic Packaging and Production Conference-Proceedings of the Technical Program (West and East)* (pp. 1423-1430). Norwalk: Reed Exhibition Companies.

F. Helo, K. P. (2000). Methodology for Predicting Manufacturing Yield for Printed Circuit Board Assemby Lines. *Journal of Electronics Manufacturing , 10* (2).

G. G. Vining, a. S. (2005). *Statistical Methods for Engineers.* Duxbury Press.

Madge, R. (2005, May-June). New Test Paradigms for Yield and Manufacturability. (Y. Zorian, Ed.) *IEEE Design & Test of Computers* , pp. 240-246.

Millman, S. D. (1993). Improving Quality: Yield Vs. Test Coverage. *International Conference on Wafer Scale Integration* , 279-288.

Patel, N. R. (2003). Lecture Notes for Course 15.062 Data Mining. *MIT Open Courseware* . Cambridge, MA: MIT http://ocw.mit.edu/

Shina, S. G. (2002). *Six Sigma for Electronics Design and Manufacturing.* McGraw-Hill.

T. Chen, V.-K. K. (1999). IC Yield Estimation at Early Stages of the Design Cycle. *Microelectronics Journal , 30*, 725-732.

Y. Li, R. L. (1994). Design Factors and Their Effect on PCB Assembly Yield - Statistical and Neural Network Predictive Models. *IEEE Transactions on Components, Packaging, and Manufacturing , 17* (2).